Priority-based Flow Control: Proposal for PAR & 5 Criteria

19 July 2007

Introduction

- Objective
  - Enable Ethernet to provide flow control characteristics similar to that of HPC and storage fabrics (e.g., Fibre Channel and InfiniBand)
    - Reduce frame loss to approx link BER (10^{-12} or better)
    - This along with Congestion Management eliminates need for end nodes to deal with congestion at higher levels (back-off, slow restart, etc.)
  - Allow conventional traffic to co-exist on such fabrics
    - Not all traffic is flow controlled
    - Segregated by priority code points

Why is this needed?

- For example: Fibre Channel over Ethernet
  - Huge market opportunity
  - Many companies aggressively pursuing
- But, Fibre Channel does not expect frames to be lost due to congestion
  - Large transfers typical (2 MB, for example)
  - Loss frame results in entire transaction to be retried
    - Sometimes at hardware / microcode level
    - More congestion -> more retries -> congestion collapse
- Frame loss rate is not the issue
  - Fibre Channel Protocol does not expect frame loss due to congestion
    - Therefore, does not respond appropriately when it occurs
- Similar arguments could be made for other HPC protocols

Where is it used?

- Used only in constrained short-range networks (similar to congestion management) with extents typically found in storage or HPC networks
  - Flow controlled networks have broad market adoption in these cases
- Like congestion management, there is no intention to extend this to larger topologies
  - Flow control and Congestion Spreading:
    - Per-Priority – restricts to only relevant traffic classes, does not eliminate congestion spreading
    - Congestion Notification addresses spreading by slowing actual sources

Title (2.1)

- Amendment to 802.1Q
- Standard for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks – Amendment 9: Priority-based Flow Control.

PAR Scope (5.2)

- This standard specifies protocols, procedures and managed objects that support flow control per VLAN tag encoded priority code points within network domains of limited bandwidth delay product. This mechanism is intended to eliminate loss due to congestion. This is achieved by a pause mechanism similar to the Ethernet PAUSE, but operating on individual VLAN tag encoded priority code points. This mechanism enhances support for and co-existence of higher layer protocols that are highly loss sensitive. VLAN tag encoded priority code points are allocated to segregate frames subject to flow control, allowing simultaneous support of both flow controlled and other higher layer protocols.

PAR Scope (5.3)

- Is the completion of this document contingent upon the completion of another document? No

PAR Purpose (5.4)

- Data center networks employ higher level protocols that depend on the delivery of data frames with a much lower probability of frame loss than is typical of IEEE 802 VLAN bridged networks. This amendment will support the use of these applications as well as their co-existence with traditional LAN applications on a single VLAN bridged network.

Need for the project (5.5)

- There is significant customer interest and market opportunity for Ethernet as a consolidated Layer 2 solution in high-speed short-range networks such as data centers, backplane fabrics, single and multi-chassis interconnects, computing clusters, and storage networks. These environments currently use Layer 2 networks that offer very low frame loss (e.g., FibreChannel, InfiniBand). This project will bring to Ethernet frame loss characteristics comparable to the ones provided by the Layer 2 networks that are currently used in these environments. Use of a consolidated network will realize operational and equipment cost benefits.

Stakeholders for the Standard (5.6)

- Developers and users of networking for data center environments including networking IC developers, switch and NIC vendors, and users.

Five Criteria

Broad Market Potential

a) Broad sets of applicability

- Mechanisms to avoid frame loss due to congestion are essential to support the highly loss sensitive higher layer protocols used for data storage, clustering, and backplane fabrics. Back-end data storage networks, clustering networks and backplane fabrics are typically limited in size, making them amenable to a flow control mechanism that operates hop by hop.
- The data traffic to be controlled by the proposed flow control mechanism will be segregated using the VLAN tag encoded priority code points, ensuring that traffic types that are not amenable to flow control may co-exist with those that are.

b) Multiple vendors and numerous users

- Multiple equipment vendors have expressed interest in the proposed project. There is strong and continued user interest in converting existing networks to Ethernet and in the realization of operational and equipment cost savings through use of a consolidated network. There is strong interest in increased use of data storage networks, provided that they can be realized with familiar technologies over a consolidated network.

c) Balanced costs (LAN versus attached stations)

- The introduction of this flow control mechanism is not expected to materially alter the balance of costs between end stations and bridges. Significant equipment and operational costs savings are expected as compared to the use of separate networks for traditional LAN connectivity and for loss/latency sensitive applications.

Compatibility

- The proposed standard will be an amendment to 802.1Q, and will interoperate and coexist with all prior revisions and amendments of the 802.1Q standard.
- The data traffic to be controlled by the proposed flow control mechanism will be segregated using VLAN tag encoded priority code points, thus ensuring that traffic types already supported by VLAN Bridges are not affected.
- The proposed amendment will contain MIB modules, or additions to existing MIB modules, to provide management operations for any configuration required together with performance monitoring for both end stations and bridges.

Distinct Identity

a) Substantially different from other IEEE 802 standards.

IEEE Std 802.1Q is the sole and authoritative specification for priority aware Bridges and their participation in LAN protocols. No other IEEE 802 standard addresses priority based flow control by bridges.

b) One unique solution per problem (not two solutions to a problem)

The need to subject certain classes of traffic to flow control mechanisms while allowing others to operate without has not been anticipated by any other IEEE802 specification; consequently, this proposal is the only solution to the problem of allowing a coexistence of such traffic types.

c) Easy for the document reader to select the relevant specification.

IEEE Std 802.1Q is the natural reference for priority based handling of traffic flows, which will make the capabilities added by this amendment easy to locate.

Technical Feasibility

a) Demonstrated system feasibility.

Similar techniques are widely deployed in other networking technologies, such as Fibre Channel and InfiniBand. The proposal is a natural extension of the expedited forwarding capability defined in IEEE Std. 802.1Q and widely deployed in bridge products.

b) Proven technology, reasonable testing.

These and similar techniques have been proven in real world deployments of Fibre Channel, InfiniBand, and other networking technologies. These techniques have been shown to be reasonably testable.

c) Confidence in reliability.

These and similar techniques have been proven reliable in real-world deployments of Fibre Channel, InfiniBand, and other networking technologies.

d) Coexistence of 802 wireless standards specifying devices for unlicensed operation.

Not applicable.

Economic Feasibility

a) Known cost factors, reliable data.

The proposed amendment will retain existing cost characteristics of bridges including simplicity of queue structures and will not require maintenance of additional queues or queue state beyond the existing per traffic class (priority) queues for conformance to either its mandatory or optional provisions. In particular per flow queuing will not be required.

The proposed amendment may require some functions, specifically the generation of per-priority flow control frames, at a rate and within a time not practical for some existing and otherwise conformant bridge implementation architectures. However these functions can be performed by some existing bridges with known implementation costs.

b) Reasonable cost for performance.

The proposed technology will reduce overall costs where separate networks are currently required by enabling the use of a consolidated network. The proposed solution allows a network to avoid frame loss due to congestion without significant throughput reduction.

c) Consideration of installation costs.

Installation costs of VLAN Bridges or end stations are not expected to be significantly affected; any increase in network costs is expected to be more than offset by a reduction in the number of separate networks required.

Input from .1 plenary

- Remove Ethernet
- Full-duplex
- Title “Data Center bridging”?
- Tie to CN?
- Clarify that it isn’t for storage/HPC across the WAN